

Informed selection of future climates

Channing Arndt • Charles Fant • Sherman Robinson •
Kenneth Strzepek

Received: 18 March 2013 / Accepted: 22 May 2014 / Published online: 19 July 2014
© UNU-WIDER 2014

Abstract Analysis of climate change is often computationally burdensome. Here, we present an approach for intelligently selecting a sample of climates from a population of 6800 climates designed to represent the full distribution of likely climate outcomes out to 2050 for the Zambeze River Valley. Philosophically, our approach draws upon information theory. Technically, our approach draws upon the numerical integration literature and recent applications of Gaussian quadrature sampling. In our approach, future climates in the Zambeze River Valley are summarized in 12 variables. Weighted Gaussian quadrature samples containing approximately 400 climates are then obtained using the information from these 12 variables. Specifically, the moments of the 12 summary variables in the samples, out to order three, are obliged to equal (or be close to) the moments of the population of 6800 climates. Runoff in the Zambeze River Valley is then estimated for 2026 to 2050 using the CliRun model for all 6800 climates. It is then straightforward to compare the properties of various subsamples. Based on a root of mean square error (RMSE) criteria, the Gaussian quadrature samples substantially outperform random samples of the same size in the prediction of annual average runoff from 2026 to 2050. Relative to random samples, Gaussian quadrature samples tend to perform best when climate change effects are stronger. We conclude that, when properly employed, Gaussian quadrature samples provide an efficient and tractable way to treat climate uncertainty in biophysical and economic models.

This article is part of a Special Issue on “Climate Change and the Zambezi River Valley” edited by Finn Tarp, James Juana, and Philip Ward

C. Arndt (✉)

World Institute for Development Economics Research, United Nations University, Helsinki,
Katajanokanlaituri 6 B FI-00160, Finland
e-mail: channing@wider.unu.edu

C. Fant

Center for Climate and Civil System, University of Colorado, Boulder, CO 80309, USA

S. Robinson

International Food Policy Research Institute, Washington, DC, USA

K. Strzepek

Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology,
Cambridge, MA, USA

1 Introduction

Socioeconomic analysis of the impacts of climate change has often proceeded on the basis of a limited number of future climate scenarios. For example, the World Bank's study of the economics of adaptation to climate change (World Bank 2011) focused on four future climates. Recent work in Vietnam expanded the number of climates considered by analyzing all 56 climates employed for the Fourth Assessment Report (AR4) (Arndt et al. 2012b). However, while these 56 climates exhibit significant variation, especially at country level scales, they do not represent a systematic attempt to populate the future distribution of potential climate outcomes for a given global climate policy scenario. Schlosser and Strzepek (2013) address this problem by generating hybrid frequency distributions (HFDs) for the Zambeze River Valley. These HFDs represent the best current attempt to describe the potential distribution of future climate outcomes out to 2050 for a particular region.

While the HFDs provide an explicit discrete estimation of the distribution of future climates, they also present other practical problems. In particular, the climate distribution for the Zambeze River Valley is populated by 6800 distinct future climates. This is an inconveniently large number for socioeconomic analysis for purely computational reasons. For example, CliCrop is a slimmed-down crop model designed specifically for the analysis of climate change issues (Fant et al. 2012). Nevertheless, solving CliCrop for multiple crops, multiple sub-regions and multiple time periods for all 6800 climates would take weeks on a single processor. Output from CliCrop then serves as input to economy-wide economic models, which also have significant computational demands. And, this exercise has to be done for each set of 6800 climates corresponding to a particular global mitigation policy scenario. Finally, the ability of simulation models to decompose and thus help explain impacts forms a significant part of their value. However, decomposition requires running the models with various subsets of shocks imposed, which significantly increases the number of times that the full climate distribution would need to be run.

In short, there are significant practical advantages to working with a more manageable subset of climates that represents the full distribution. In general terms, this is hardly a new problem. The impracticalities of obtaining information from an entire population are the genesis of sampling theory. Indeed, one defensible way to proceed with the problem at hand would be to choose a random subset of J climates from the population of 6800 and restrict analysis to only those J climates. However, in this case, simple random sampling may well represent an unnecessarily uninformed approach to selecting climates. In addition, in random sampling, there is positive probability that a distinctly unrepresentative sample will be selected.

Our proposed approach to selecting climates is motivated, philosophically, by information theory (Shannon 1948; Judge and Mittelhammer 2012), which dictates that one should make complete use of available information while taking care not to impose any additional assumptions (information) either explicitly or implicitly. In our case, we know, for example, that persistent drying is highly likely to lead to reduced runoff, reduced hydropower output, and, for most crops, reduced yield. This knowledge opens the possibility of selecting climates on the basis of important aspects of the distribution of future climates. Technically, our approach follows the lead of Arndt, Kozlitina, and Preckel (2006) who employ Gaussian quadrature, a technique from the numerical integration literature, to the problem of sampling in the presence of information about the behavior of the target variables of interest.

In short, we postulate that salient aspects of the distribution of future climates can be summarized via the calculation of a limited number (12 in our case) of summary variables that describe the climate distribution. We then select a quadrature. A standard Gaussian quadrature of order three would contain a weighted set of climates where, for the 12 summary variables in

question, the moments of the distribution of the sample equal the moments of the parent distribution out to order three.¹ Hence, the quadrature (i.e., the sample) and the parent distribution share the same moments (e.g., mean, variance, and skewness terms) for the identified 12 summary variables. If there exists an order three polynomial in the 12 summary variables that provides a good approximation to the outcomes of interest, then the selected sample will provide a good approximation to the moments of the outcome variables.

We test this proposition in the Zambeze River Valley using the HFDs mentioned above and the CliRun model of runoff (Strzepek et al. 2011). CliRun is practical to solve for the entire set of HFDs corresponding to two future scenarios: (i) unconstrained emissions (UCE) where no policy actions are taken to limit greenhouse gas emissions out to 2050 and (ii) restraints on global emissions to prevent global greenhouse gas concentrations from exceeding 560 ppm CO₂ equivalent, which is labeled ‘level one stabilization’ (L1S) in Webster et al. (2011). For each scenario, we compare 100 Gaussian quadratures and 100 random samples of the same size with the ‘true’ distribution of runoff results (i.e., the distribution derived from running all 6800 climates through CliRun). We focus on the change in annual average runoff in the eastern and western Zambeze due to climate change. We find that the quadratures perform well, particularly in the UCE scenario and in later periods (the 2040s) when climate change impacts are more pronounced.

The remainder of this paper is structured as follows. Section 2 briefly reviews the theory underlying the proposed approach. Section 3 presents the actual procedures employed for selecting future climates. Section 4 considers the performance of these samples relative to random samples of similar size. Section 5 concludes that the approach described here, judiciously implemented, holds considerable promise. Section 5 also provides directions for future research.

2 Theory and methods

2.1 Gaussian quadrature

In general, the task of calculating the moments of the distribution of the outputs of a process (denoted as $f(x)$) with random inputs drawn from the distribution $g(x)$ over the domain Ω can be viewed as a problem of integration as shown in equation (1). Frequently, an analytical solution to the right hand side of equation (1) does not exist. In these cases, numerical approximations become necessary. These approximations are often a weighted sum of evaluations of the integrand as shown on the left hand side of equation (1)

$$\sum_{j=1}^J w_j f(x^j) \approx E[f(x)] = \int_{\Omega} f(x)g(x)dx \quad (1)$$

where the vector x^j represents the points where $f(x)$ is evaluated and w_j are the weights associated with each evaluation. Various approaches exist to choose the points and weights.

One possible approach to selecting points and weights is to randomly choose a large number of points, L , from the distribution $g(x)$ within the domain Ω and assign equal weight, $w_j = 1/L$, to each point. If L is sufficiently large, the method will produce a close approximation to the integral under very general conditions. Due to these properties—and modern

¹ We focus on Gaussian quadratures of order three for practical reasons that are clarified later, particularly in equation (3).

computational capacity—this Monte Carlo approach is very attractive in many instances. However, when evaluating the function $f(x)$ is computationally burdensome, reducing the number of required evaluations, without substantial loss in accuracy, becomes attractive.

Haber (1970) reviews formulae for finding accurate and efficient approximations to definite integrals. One set of formulae, called Gaussian quadratures, are especially appealing due to their computational efficiency under broad conditions. Gaussian quadratures efficiently estimate definite integrals when the integrand is well approximated by a polynomial in x of sufficiently high degree, d . By choosing a polynomial of sufficiently high degree, any smooth, bounded integrand can be well approximated. Usefully, if the integrand is, in fact, a polynomial of order d or less, then an approximation to the definite integral based on a Gaussian quadrature of order d or more will be exact.

To obtain a Gaussian quadrature of order d , one must solve the following system of equations:

$$\sum_{j=1}^J w_j \prod_{i=1}^M (x_i^j)^{l_i} = \int_{\Omega} \left[\prod_{i=1}^M (x_i)^{l_i} \right] g(x) dx \quad (2)$$

where $l_i \in (0, 1, 2, \dots, d)$, such that $\sum_{i=1}^M l_i \leq d$, and M is the dimension of the vector x .

The right-hand sides of the above set of equations represent the moments of x about the origin. The total number of equations in the system will be equal to the number of unique moments up to order d :

$$J = \binom{M+d}{d} = \frac{(M+d)(M+d-1)\dots(M+1)}{d(d-1)\dots(1)} \quad (3)$$

For instance, if we have 12 variables and draw an order three quadrature, the number of equations in the system will be 455.

Tchakaloff (1957) proves that this system can always be solved with no more than J points, all of which have positive weight and lie within the domain of integration. In general, there are multiple solutions to the system of equations in (2) implying that Gaussian quadratures are not unique. From (3), it is clear that, as the dimensionality of integration, M , and/or the order of approximation, d , expands, the upper bound on the number of points in the quadrature, J , increases rapidly. For this reason, Gaussian quadrature may not apply practically for integration problems of high dimensionality or for problems with integrands that demand a high order polynomial for adequate approximation.² Nevertheless, empirical experience with integration problems of dimension 12 or less with smooth integrands over a closed domain of integration suggests substantial efficiency gains even when the order of approximation is relatively low ($d=3$) (DeVuyst and Preckel 1997).

2.2 Gaussian quadrature climate sampling

The problem at hand maps reasonably well to the theoretical development presented above. In our case, the integrand is the CliRun model and x contains random climate variables as derived from the HFDs. There are two departures from straightforward numerical integration via Gaussian quadrature as described above. First, our case pertains to discrete distributions rather than the continuous distributions employed above. Given the number of climates in the

² With dimension 12 and an order four quadrature ($d=4$), the maximum number of points required grows to 1820.

parent HFD, this departure is not particularly problematic. Second, and more importantly, the random vector of climate outcomes x entering CliRun is of massive dimensionality (M). From equation (3), the straightforward application of Gaussian quadrature is not practical due to the number of points required.

To surmount this difficulty, we hypothesize that substantial information inherent in the distribution of climate outcomes can be summarized in a limited number of variables, z . For example, maximum monthly precipitation in a given year may be a good indicator of flood probability in that year. Note that the variables z are highly likely to be functions of x . Once these variables [$z = s(x)$] have been identified, it is then straightforward to obtain a Gaussian quadrature of order d by solving the system of equations in (2) on the basis of the moments of z (DeVuyst and Preckel 2007). By definition, the moments of the summary variables z from the sample distribution will equal the moments of the population distribution up to order d . However, the Gaussian quadrature on z will not perfectly represent the moments of x . In this respect, the quality of approximation of the moments of the final outcome variables (the outputs of CliRun) depend upon the strength of relationship between the moments of z and the moments of x .

Following Arndt, Kozlitina, and Preckel (2006), the issue can be posed in a manner more amenable to investigation. Assume that there exists a function $h(z)$ that approximates CliRun [$f(x)$] with error:

$$y = f(x) = h(z) + e = h(s(x)) + e \quad (4)$$

If we assume, for a discrete population of size N , that the expected value of the error term is approximately zero and that all points are equally likely, then the mean of y is equal to:

$$\mu^y = \frac{1}{N} \sum_{n=1}^N y^n = \frac{1}{N} \sum_{n=1}^N (h(z^n) + e^n) \approx \frac{1}{N} \sum_{n=1}^N h(z^n) = E[h(z)] = E[h(s(x))] \quad (5)$$

Assuming that $h(z)$ is in fact a polynomial of order d or less, a Gaussian quadrature of order d (or more) will exactly approximate the right-hand side of equation (5).³ However, since we do not observe $h(z)$ the Gaussian quadrature approximation to the mean of y must rely on $f(x)$ and hence includes a systematic component $h(z)$ and an error term:

$$\bar{y}_{gq} = \sum_{j=1}^J w_j y^j = \sum_{j=1}^J w_j f(x^j) = \sum_{j=1}^J w_j (h(z^j) + e^j) \quad (6)$$

From equation (6), it is clear that the accuracy of the quadrature-based estimate of the mean of y depends in significant measure on the quality of approximation of $f(x)$ by a polynomial of order d $h(z)$. In the (highly unlikely) event that $h(z)$ exactly approximates $f(x)$ within the relevant domain, Ω , then the Gaussian quadrature approximation to the mean of y will be exact. This is obviously superior to a random sample of climates. If, on the other hand, the relationship between $h(z)$ and $f(x)$ within the domain Ω is weak, then the properties of the Gaussian quadrature sample are difficult to discern and are not likely to be appealing. In addition, unequal weights are effectively applied to the error term in equation (6), likely increasing the variance of the estimate relative to a random sample of similar size.

The wisdom of the application of the approach then depends substantially upon the quality of approximation of $f(x)$ by $h(z)$. A detailed measure of the quality of this approximation is the

³ Henceforth, we refer to z and $h(z)$ rather than to $z=s(x)$ and $h(s(x))$ in order to reduce notational clutter.

R-squared from a standard least squares regression procedure of y on an order d polynomial in z . For instance, for an order three quadrature, the regression equation would take the following form:

$$y = \delta + \sum_{i=1}^M \beta_i z_i + \sum_{i=1}^M \sum_{j=1}^M \alpha_{ij} z_i z_j + \sum_{i=1}^M \sum_{j=1}^M \sum_{k=1}^M \gamma_{ijk} z_i z_j z_k + \xi \quad (7)$$

where $i \leq j \leq k$. It merits emphasizing that (7) is not useful in practice as the purpose of the approach presented here is to avoid calculating y for every point in the population. We do so here in order to investigate the properties of the approach.

We turn now to an application.

3 Runoff in the zambeze river Valley under climate change

The primary climate variables needed for the subsequent climate impact modeling, derived from the two HFD scenarios, are near-surface temperature and precipitation. As shown in Schlosser and Strzepek (2013), the UCE scenario presents a warmer future climate than the L1S policy case. In terms of precipitation, there are only subtle differences for the median prediction in the two policies with a higher chance of an increase. The more defining characteristic of the predicted climate in the UCE policy case compared to L1S is the distribution shape, where the UCE predictions show a much wider distribution. This characteristic is especially noticeable in the tails of the predicted precipitation change, implying that UCE policies will result in a higher possibility of extreme climate differences by 2050 increasing the uncertainty of the future.

These temperature and precipitation outcomes serve as inputs into CliRun, which is a rainfall runoff hydrologic model designed specifically for understanding the impact of changes in climate on surface water availability. Based on the set of equations solved in CliRun, the model is nonlinear, continuous, and differentiable. CliRun was calibrated using runoff data generated by the University of New Hampshire in collaboration with the Global Runoff Data Centre (UNH-GRDC; Fekete, Vörösmarty, and Grabs 2002). Changes in climate from the full HFD climate pool were applied to the historical monthly mean climate over selected representative basins in the Greater Zambeze River Valley. Distributions of change in runoff for the east and west regions illustrate the same patterns as the changes in climate discussed above, where UCE presents a distribution with greater variance than L1S.

4 Selecting future climates

4.1 Developing climate summary variables

Using climate statistics that best explain the resulting climate impacts, important for the subsequent modeling, is an essential step in the process. These summary variables define the differences and similarities between the various climate scenarios, guiding the final selection of climates that will enter the impact modeling. So, the variables chosen need to be descriptive in terms of the impact modeling efforts, relatively simple to calculate and few in number. In addition, the intent is to select a subset of climates for use in crop, flooding, infrastructure, and economic models as well as for the CliRun model. Hence, the variables are

chosen to reflect the diversity of likely impacts of climate change and not exclusively for the purposes of approximating runoff output from CliRun.

First, the entire region was split into two sub-regions, east and west, in order to capture some of the geographically significant differences. Second, six climate variables were chosen in each region, making 12 summary variables overall. The first three variables were maximum monthly precipitation in 2050, and maximum monthly change in temperature for both 2030 and 2050. These variables were selected to take into account the impacts of climate on infrastructure, where high temperatures and heavy rainfall can greatly compromise the integrity of roads and buildings. The next two variables chosen in each region were the climate moisture index (CMI) for both 2030 and 2050. The CMI is a simple measure of annual aridity depending on total annual precipitation and potential evapotranspiration directly. CMI ranges from -1 to $+1$, where -1 is very dry and $+1$ is very humid. This variable was used as a simple measure to indicate changes in soil moisture and surface water availability important for the water resource modeling efforts. Changes in CMI are also important for changes in agricultural yield, irrigation water availability, and hydropower generation. The sixth variable chosen is the standard deviation of the change in seasonal precipitation in 2050. In the Zambeze basin, this value would indicate a shift in the growing season, which farmers attempt to line up with the rainy season in such a way as to obtain the highest yield. A significant shift in the rainy season could greatly impact crop production and irrigation demand.

In all, eight summary variables are employed from 2050 (four each for East and West) and four summary variables are employed from 2030. The later period, 2050, receives more weight since it is the focus of the analysis. Nevertheless, four variables from 2030 are included because many climate change impacts, particularly economic impacts, reflect long run path dependent processes. Consequently, measures of the onset of climate change in the relatively near term (2030) are important.

4.2 Operational procedures and rationale for additional constraints

Having selected 12 variables z which are intended to summarize the complete distribution of the climate variables x , it is straightforward to calculate the moments of z up to order d (mean, variance, skewness, and all cross terms). Not coincidentally, the number of moments is equal to the Tchakaloff bound, which includes moment zero (the sum of the weights must equal one). From equation (3), we see that, with 12 variables and $d=3$, the upper bound on the number of points in a Gaussian quadrature is 455. Alternatively viewed, equation set (2) contains 455 equations.

Note that equation set (2) is linear in the weights, w . Hence, equation set (2) can form the constraint set of a linear program with 6800 columns (equal to the number of HFD climates) and 455 rows. Further, note that the linear program is feasible. By construction, the system will be satisfied when all weights are set at $1/6800$. Algorithms that produce extreme point solutions, such as the simplex method, will solve the system with a maximum of 455 strictly positive weights as long as starting values for all weights are set at zero (DeVuyst and Preckel 2007).

If only one quadrature is desired, then a degenerate objective, such as $obj=1$, can be employed. In our case, we wish to draw multiple quadratures (recall that they are not unique) in order to compare the properties of the quadratures with the properties of a series of random samples. To accomplish this, we minimize a standard linear programming objective equal to the sum of weights multiplied by a coefficient ($obj = \sum_{i=1}^{6800} C_j * W_j$). For the first quadrature, the values of the coefficients, c , are set equal to one (note that this objective replicates the moment zero constraint; hence the value of the objective will be equal to one). In order to obtain the

Table 1 Average ratio (Gaussian quadrature/random samples) of root of mean square error, 2026–50

		Unconstrained emissions		LIS	
		Moment 1	Moment 2	Moment 1	Moment 2
Moment 1 corresponds to mean runoff anomalies and moment 2 to the variance of runoff anomalies about zero	West	0.740	0.764	0.931	0.991
	East	0.633	0.629	0.888	0.938
	Total	0.687	0.696	0.909	0.964

second objective, the coefficient on each weight that was strictly positive (e.g., included in the first quadrature) is incremented by one. Hence, for all climates included in the first quadrature, the corresponding coefficient, c , on the weight, w , in the objective function will be set at two. This process is then repeated. So, if a climate (also called a point in the quadrature) has appeared in both the first and second quadratures, it will have a coefficient of three. If the point has appeared in either quadrature one or two, it will have a value of two. If it has not appeared in either of the first two quadratures, it will have value one. Minimizing this objective encourages the routine to choose different climates to enter the quadrature while still satisfying the constraint set. This procedure results in 100 distinct quadratures.⁴

For the problem at hand, some additional constraints are applied. As explained in Schlosser and Strzpek (2013), the 6800 climates are derived from 400 runs of the integrated global systems model combined with the regional results for the Zambeze river valley from 17 general circulation models (GCMs). We add additional constraints that assure that the mean, variance, and skewness (excluding covariance and cross skewness terms) is met for each GCM for four of the 12 summary variables (the four related to precipitation). In addition, we place upper and lower bounds on the weight associated with each GCM. As a result of these constraints, all quadratures will be selected based on a range of GCMs, which is desirable. At the same time, these additions represent 204 ($=4*17*3$) additional equality constraints and 34 inequality constraints to the system, which increases the number of points required in the quadrature. To reduce the number of points, we relax the conditions on skewness and cross skewness terms in equation set (2). The quadratures must generate skewness and cross skewness terms within 1% of the value for the full population.⁵

While the Tchakaloff bound provides an upper limit on the number of points in the quadrature, quadratures can often be found with fewer points. For the problem at hand, quadratures contained between 363 and 432 points with a mean of 411 for the UCE scenario and between 341 and 409 points with a mean of 389 for the LIS scenario. We turn now to how well these quadratures predict key outputs from CliRun.

5 Performance of the gaussian quadrature samples

5.1 Comparison versus random samples

The performance of the Gaussian quadrature approach is verified through a comparison with random samples. As discussed above, the number of points in each of the 100 quadratures can

⁴ This was verified by observing that the sum of the squares of the weights for each quadrature is unique.

⁵ Loosening of constraints on moment three conditions merits further research.

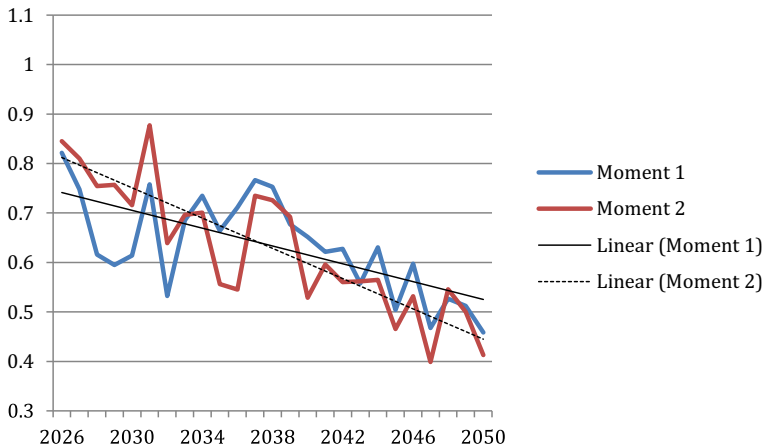


Fig. 1 Ratios of root of mean square error (vertical axis) for Gaussian quadrature/RS in the east for the UCE scenario through time

vary. To account for this, random samples are drawn of size equal to the corresponding quadrature. In other words, if the fifth quadrature obtained has size 410, then the fifth random sample drawn will also have size 410. In addition, as discussed in Section 4.2, various constraints were added to the linear programming problem to guarantee that each of the 17 GCMs used to develop the climate HFDs are represented. Similarly, stratified random samples were drawn across the 17 GCMs.

To compare the random samples with the Gaussian quadrature samples, we focus on predicted annual runoff anomalies for the eastern and western Zambeze for the years 2026–50 treating the average anomaly across the 6800 climates as the population mean. We calculate the average RMSE of the first and second moments of climate change induced anomalies in runoff in the eastern and western Zambeze for each year from 2026–50 across the 100 Gaussian quadrature and random samples. To facilitate comparison, we take the ratio of these two average RMSEs with the Gaussian quadrature sample in the numerator and the random sample in the denominator. Hence, ratio values less than one indicate superior performance of the Gaussian quadrature sample.

Table 1 illustrates the simple average of the ratio across all years (2026–50). In all instances, Gaussian quadrature samples outperform random samples on average. Gaussian quadrature samples perform better relative to random samples under the UCE scenario compared with the LIS scenario. In addition, runoff in the eastern Zambeze is more accurately predicted, relative to random samples, than runoff in the western Zambeze. Overall, efficiency gains are particularly strong in the UCE scenario and in the east. Viewed in another way, random sample size would have to approximately double in order to attain the efficiency levels of the quadratures in the UCE scenario for the east.⁶ This implies a doubling of the overall computational burden for

⁶ This is calculated by taking the ratio of the deflation factors for sample means: $\frac{\left(\frac{1}{\sqrt{n}}\right)}{\left(\frac{1}{\sqrt{2n}}\right)} = 0.7$ where the right hand side value of 0.7 is slightly more than the RMSE ratios obtained in Table 1 for the UCE scenario in the east.

Table 2 R-squared of regressions of y on z and standard deviation of y

	R-squared		Standard deviation	
	UE	L1S	UE	L1S
West	0.80	0.78	8.43	6.84
East	0.85	0.75	8.47	6.81
Total	0.83	0.76	8.48	6.74

The standard deviation of historical runoff is approximately 15

random samples in order to obtain the same level of efficiency. For Level 1 Stabilization, the quadratures produce gains, but these gains are of a lesser magnitude and are quite small for moment 2 in the west.⁷

Further insight can be obtained by examining the time path of the ratios. A representative time path, from the eastern region under the UCE scenario, is shown in Fig. 1. A number of interesting additional observations emerge, which also apply to time paths for the western region and the L1S scenario. First, in all four cases (East-UCE; West-UCE; East-L1S; and West-L1S), the performance of the quadratures improves with time.⁸ By 2050, the quadratures are outperforming random samples by 10–55%. Second, the trend improvement in moment 2 is more rapid in all instances. Third, there is a tendency for the quadratures to do well relative to the linear trend line around 2030 (e.g., below the trend line), relatively poorly in the late 2030s, and relatively well in the late 2040s. Recall that the vector z contained variables summarizing the period around 2030 and 2050. Hence, the quadratures appear to be doing well nearer to the periods referred to in the vector z .

As discussed in Section 2, regression analysis whereby the target runoff variables are regressed against the z vector using the full sample can provide some insight into these results. Performance of the quadratures relative to random samples also depends upon the performance of the random sample. *Ceteris paribus*, the expected RMSE of a random sample will increase as the standard deviation of the target variable, y , increases. With Gaussian quadrature, two factors are at work: the approximation of systematic variation as captured by $h(z)$ which is very efficient, and the estimation of the error term, e , which is inefficient due to the presence of unequal weighting (Arndt, Kozlitina, and Preckel 2006).

Table 2 shows the average R-squared of linear regressions of runoff anomalies, y , on the summary variables, z , as well as the corresponding average standard deviation of the runoff anomalies. The regressions are run (and standard deviations of runoff anomalies calculated) for each year from 2026–50 and for the eastern and western zones of the Zambeze. The averages are taken over time. The table provides some insight into the results obtained. When comparing the UCE scenario with the L1S scenario, one notes that the R-squared values and standard deviation values are higher in the UCE scenario. This helps to explain the relatively good performance of the Gaussian quadrature sampling approach in the UCE scenario. The high R-squared provides a boost to the Gaussian quadrature approach while the relatively high standard deviation of the target variable decreases the performance of random samples.

⁷ Mean percentage errors are approximately zero for both the quadratures and (as expected) the random samples.

⁸ Graphs for all four cases can be found in Arndt et al. (2012a).

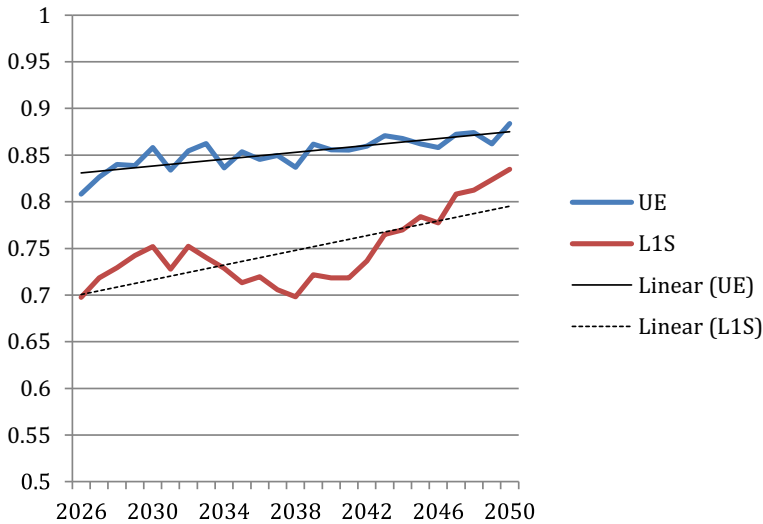


Fig. 2 R-squared of regression of runoff on z for the eastern Zambeze (vertical axis) through time

The relatively strong performance of Gaussian quadrature samples in the east compared with the west is also consistent with these explanations. Runoff in the east is both better predicted (higher R-squared) and more variable (higher standard deviation) in the UCE scenario. For L1S, the R-squared measures favor the west while the standard deviation measures favor the east.

Similar to the ratios shown in Fig. 1, examining the R-squared values through time provides further insight. Representative R-squared values for the east are presented in Fig. 2. For both the east and the west, R-squared values trend upward with time. One can also note that R-squared relative to trend is higher around 2030, lower around the late 2030s, and higher towards the end of the first half of the 21st century, which is a mirror image of the tendency for the RMSE ratios. The upward trend in R-squared through time combined with a trend increase in the standard deviation of runoff (not shown) help to explain the performance improvement of Gaussian quadrature sampling through time as shown in Fig. 1.⁹

5.2 Selecting an optimal quadrature

For the purposes of assessing the performance of Gaussian quadrature sampling, we have run the CliRun model for all 6800 climates and drawn multiple quadratures and multiple stratified random samples of the same size. This permits us to compare the relative performance using a RMSE criteria. However, in real applications of Gaussian quadrature sampling, ‘truth’ will not be known as we are expressly attempting to avoid running a chosen biophysical or economic model for all climates. In addition, we wish to choose only one Gaussian quadrature sample.

⁹ A factor driving the more rapid trend increase in the relative performance of Gaussian quadrature sampling in estimating moment 2 of the runoff anomalies is a more pronounced increase in the standard deviation of the square of the runoff anomalies.

To choose the selected Gaussian quadrature sample of the 100 generated, we choose the minimum of

$$\left[\frac{\sum_{j=1}^J w_j * w_j}{\sum_{j=1}^J \frac{1}{J * J}} \right] \quad (8)$$

across all 100 samples. This corresponds to choosing the quadrature with the smallest variance inflation factor relative to a random sample of the same size. The logic for this choice is derived from equation (6), which divides Gaussian quadrature sample estimates into systematic and error components. A priori, there is no reason to prefer one quadrature over another (except with respect to size, where smaller is better) for the estimation of the systematic component. However, the Gaussian quadrature sampling estimate of the error component can be compared to random sampling with unequal weights. In the absence of heteroskedasticity, the application of unequal weights is inefficient precisely because the denominator is smaller than the numerator in equation (8). Hence, the choice minimizes this inefficiency.¹⁰

In our case, the selected Gaussian quadrature samples are of size 426 for UCE scenario and size 398 for the LIS scenario.

6 Conclusions, limitations and future research

We conclude that Gaussian quadrature sampling provides a promising approach to selecting an informed sample of future climates from a large population. Given the ubiquity of computing challenges in the climate change area, the technique quite likely has other applications. At the same time, the limitations of the Gaussian quadrature approach should also be borne firmly in mind. Without a sufficiently strong relationship between the summary variables (denoted here as z and the outcome variable of interest (denoted here as y), Gaussian quadrature samples can produce outcomes inferior to random samples. In addition, the economic and biophysical models generating the outcome variable(s) of interest should be smooth over the (closed) domain of interest.

With these caveats in mind, Gaussian quadrature sampling appears to offer considerable promise. While the mathematics of numerical integration formulae in general and Gaussian quadrature in particular is somewhat abstruse, the intuition is easily communicated. It is intuitively attractive to have the moments of a sample match the moments of the parent population. Once this concept has been explained, some simple examples often provide sufficient reinforcement (see, for example, Arndt, Kozlitzina, and Preckel (2006) for the case of univariate quadratures).

With respect to future research, focus should be placed on the choice of summary variables, z . For example, in the economics of climate change, results are normally driven by a relatively limited number of effects. These include average yield impacts weighted by the value added shares of crops, changes in hydropower production, and changes in the frequency or intensity of extreme events such as floods and cyclones. In this sense, the biophysical models that convert climate outcomes into economically meaningful impacts are prime candidates for creation of the z summary variable vector. The issue, as mentioned in the introduction, is that

¹⁰ Note that a simpler measure, such as the variance of the weights, will tend to favor quadratures with more points. This occurs because the weights are constrained to sum to one regardless of the number of points.

many biophysical models are themselves computationally burdensome to solve. A potential solution lies in the development/use of simplified biophysical models that are informative with respect to the sign and magnitude of climate change impacts but are less refined, and hence faster to solve, than their more detailed cousins. These simplified models could be used to develop the z vector.

With future climates summarized in terms of economically relevant impacts, a highly informed subsample of climates could be selected from the parent distribution. With this much more limited number of climates in hand, analysis could proceed by running the climate sample through the best possible biophysical and economic models in order to develop the best possible estimation of climate change impacts for each chosen climate scenario. This approach likely presents the most promising way forward for the problem at hand.

References

- Arndt C, Kozlitina J, Preckel PV (2006) Efficient sub-sampling using Gaussian quadrature. *Applied statistics*. J Roy Stat Soc C-App 55:355–364
- Arndt C, Fant F, Robinson S, Strzepek K (2012a) Informed selection of future climates. UNU-WIDER working paper 2012/60
- Arndt C, Tarp F, Thurlow J (2012b) The economic costs of climate change: a multi-sector impact assessment for Vietnam. UNU-WIDER Working Paper 2012/82.
- DeVuyst EA, Preckel PV (1997) Sensitivity analysis revisited: a quadrature-based approach. *J Policy Model* 19(2):175–185
- DeVuyst EA, Preckel PV (2007) Gaussian cubature: A practitioners guide. *Math Comput Model* 45:787–794
- Fant C, Gueneau A, Strzepek K, Awadalla S, Farmer W, Blanc E, Schlosser CA (2012) CliCrop: a crop water-stress and irrigation demand model for an integrated global assessment modeling approach. Joint Program on the Science and Policy of Global Change, Report 214, MIT. <http://globalchange.mit.edu/research/publications/2264>
- Fekete BM, Vörösmarty CJ, Grabs W (2002) High resolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochem Cy* 16:3
- Haber S (1970) Numerical evaluation of multiple integrals. *SIAM Rev* 12:481–525
- Judge GG, Mittelhammer RC (2012) An information theoretic approach to econometrics. Cambridge University Press, Cambridge
- Schlosser CA, Strzepek K (2013) Regional climate change of the greater Zambezi River Basin: a hybrid assessment. UNU-WIDER Working Paper 2013/040.
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Strzepek K, McCluskey A, Boehlert B, Jacobsen M, Fant C IV (2011) Climate variability and change: a basin scale indicator approach to understanding the risk to water resources development and management. World Bank, Washington
- Tchakaloff V (1957) Formules de cubatures mécaniques à coefficients non négatifs. *Bull Sci Math* 81:123–134
- Webster M, Sokolov AP, Reilly JM, Forest CE, Paltsev S, Schlosser A, Wang C et al (2011) Analysis of climate policy targets under uncertainty. *Clim Chang* 112(3–4):1–15
- World Bank (2011) Economics of adaptation to climate change. World Bank, Washington